Why did you pick that? Capturing reasons for assigning value in exploratory search

Version: Version of Record

Link(s) to article on publisher's website:
http://dx.doi.org/doi:10.1007/s00799-011-0067-7

Why did you pick that? Visualising relevance criteria in exploratory search

Ulises Cerviño Beresi, Yunhyong Kim, Dawei Song & Ian Ruthven

International Journal on Digital Libraries

ISSN 1432-5012
Volume 11
Number 2

Int J Digit Libr (2011) 11:59-74
DOI 10.1007/s00799-011-0067-7

Springer

Int J Digit Libr (2010) 11:59–74
DOI 10.1007/s00799-011-0067-7

Why did you pick that? Visualising relevance criteria in exploratory search

Ulises Cerviño Beresi · Yunhyong Kim · Dawei Song · Ian Ruthven

Published online: 27 September 2011

Abstract In this article, we present a set of approaches in analysing data gathered during experimentation with exploratory search systems and users' acts of judging the relevance of the information retrieved by the system. We present three tools for quantitatively analysing encoded qualitative data: relevance-criteria profile, relevance-judgement complexity and session visualisation. Relevance-criteria profiles capture the prominence of each criterion usage with respect to the search sessions of individuals or selected user groups (e.g. groups defined by the users affiliations and/or level of research experience). Relevance-judgement complexity, on the other hand, reflects the number of criteria involved in a single judgment process. Finally, session visualisation brings these results together in a sequential representation of criteria usage and relevance judgements throughout a session, potentially allowing the researcher to quickly detect emerging patterns with respect to interactions, relevance criteria usage and complexity. The use of these tools is demonstrated using results from a pilot-user study that was conducted at the Robert Gordon University in 2008. We conclude by highlighting how these tools might be used to support the improvement of end-user services in digital libraries.

U. Cerviño Beresi (✉) · Y. Kim · D. Song
The Robert Gordon University, School of Computing, Aberdeen, UK
e-mail: ulises.cervino@gmail.com

Y. Kim
e-mail: yunhyong.k@gmail.com

D. Song
e-mail: d.song@rgu.ac.uk

I. Ruthven
Department of Computer and Information Sciences, The Strathclyde University, Glasgow, UK
e-mail: Ian.Ruthven@cis.strath.ac.uk

Keywords Relevance criteria · Exploratory search · Information retrieval · Literature-based discovery · User study · Document valuation

1 Introduction

Now, everyday individuals face the decision of whether or not to retain a piece of information in their personal collection, and these decisions involve a complex process of gauging the value of a document. The situation is akin to the valuation process used by an antique dealer, acurator and/or an archivist in assessing the value of an artefact: several criteria are employed to determine the object value, e.g. in terms of date, rarity, popularity and condition of the object. Likewise, the qualitative or pragmatic value of a document is determined by a number of criteria, e.g. currency, novelty, validity and clarity. The consideration of these criteria results in an overall estimate of the document's usefulness within the context of user's tasks and already accumulated information.

The criteria employed in the valuation process, although clearly related to metadata elements (e.g. date of creation) employed within libraries (e.g. Dublin Core Metadata Elements[1]), as well as the topicality of the document, do not map directly onto either of these. By studying the way in which information searchers and seekers utilise and weight these criteria, we hope to bridge the gap between human information valuation behaviour, and implementations of information retrieval (IR) engines and library end-user services.

To be able to study these criteria, one must observe their usage in a realistic scenario. The guidelines for evaluating IR systems proposed by Borlund [4] allow the researcher to gather both the system performance as well as the cognitive

[1] http://dublincore.org/documents/dces/.

data–data which includes these relevance criteria observations. Realism is achieved by involving potential end-users as test persons, and employing simulated work-task situations: descriptions of a situation in which needs for information are triggered for users. These gathered data allow the experimenter to analyse not only final results such as number of relevant objects retrieved but also the processes that led to judgements of relevance. The analysis of the performance data gathered is usually done through the examination of metrics such as precision and/or recall [6]; however analysing cognitive data such as the thought processes that led to the user-valuations of the documentation retrieved—*relevance processes* as we call them—may not be as straightforward.

In this article, qualitative data, the verbal reports of exploratory search-system users, are transformed into quantitative data using protocol analysis techniques which include transcriptions, segmentation and tagging of the segmented transcriptions. The segments are tagged with a set of relevance criteria codes, and the result is analysed using standard quantitative measures, to produce relevance profiles based on the frequency of criterion usage with respect to individuals, and groups of individuals working in similar research areas and/or having similar research experiences. We also analyse variations of complexity, that is, the number of criteria employed within a single relevance judgement.

We will further bring these together in a new visualisation technique for tracking how the relevance profiles and complexities change throughout a user session. This is implemented as colour-coded *relevance-judgement piles*, a set of relevance criteria, delimited by user interaction (e.g. navigation and reading). This approach provides a potential starting point for further study in investigating the dynamics and emerging patterns within search sessions.

The remainder of the article is structured as follows. In Sect. 2, we introduce Barry and Schamber's relevance criteria classes [2]. Section 3 describes think-aloud protocols and their processing. The main contributions of this study, namely, relevance-criteria profiles, complexity and session visualisations, are introduced and discussed in Sect. 4.1. In Sect. 5, we explore the data obtained from a user study conducted during the first half of 2008 using the techniques described in the previous sections. We conclude with some final remarks and recommendations for future study in Sect. 6.

Other studies have looked at relevance criteria in the context of information-seeking behaviour (e.g. [10,13]). This article aims to distinguish itself from these earlier studies by emphasising the need for approaches that transform qualitative results to quantitative measures on multiple levels, and by placing a special focus on the context of the literature-based discovery and exploratory search within the research domain (i.e. not for the worldwide web search).

Let us note that, while we have tried to present convincing observations in our discussion of the user study (Sect. 5), we do not claim it to be a comprehensive and conclusive study (see discussion of future work in Sect. 6). The user study is intended as a pilot study to demonstrate the potential of the tools we have introduced. To keep the focus, details regarding the study and the search engine used in the study have been kept brief (Sect. 5). Those interested in the full details of the study may, however, find them in [5].

2 Relevance criteria

Relevance judgements are often reduced to being binary judgements, or graded assessment, of relevance at best (cf. discussions in [3]) providing no explanation as to why the value was assigned. It could be that, while users consider one document to be relevant based on the length and depth of the information provided, they might consider another document relevant based on the clarity of presentation. In this article, we focus on some of the reasons that might motivate relevance judgements.

Relevance criteria are reasons expressed by users when deciding whether a given piece of information is relevant, i.e. evaluating whether to obtain and use or discard information. Barry and Schamber suggest that there is 'evidence that a finite range of [relevance] criteria exists and that these criteria are applied consistently across types of information users, problem situations, and source environments' [2]. The starting point they suggest for examining relevance criteria consists of the overlap of taxonomies resulting from two studies [1,14] on user-relevance criteria. Both studies are similar in the methodologies used; however, the types of users, information sources and formats are quite different. In this study, we extend and refine this overlap with some of the criteria appearing in Barry's original taxonomy [1] to examine 15 criteria. This selection has been motivated by the fact that the study described in Sect. 5 is almost similar to Barry's study both in terms of population and type of information searched. Although the overlap between sets would have been a good starting point, we expected some of the criteria reported by Barry [1], in particular those pertaining novelty, to be observed with regularity, e.g. document novelty. Our expectations were supported by a pilot study run before the main study (Sect. 5). The relevance criteria we have examined are listed below (the criteria added to the overlap are indicated in italics):

– Depth/scope/specificity: Whether the information is in depth or focused, has enough detail or is specific to the user's needs. Also, whether it provides a summary or overview or a sufficient variety or volume.

- Accuracy/validity: Whether the information found is accurate or valid.
- Clarity: Whether the information is presented in a clear fashion. This includes well-written documents as well as the presence of visual cues such as images.
- Currency: Whether the information is current or is up to date.
- Tangibility: Whether the information relating to tangible issues, hard data/facts are included, or information provided was proven.
- Quality of sources: Whether the quality of the information can be derived from the quality of the sources of it. This includes authors as well as publications.
- Accessibility: Whether there is some cost involved in obtaining the information.
- Availability of information/sources of information: Whether the information is available at that point in time.
- Verification: Whether other information in the field, or the user, agrees with the presented information.
- Affectiveness: Whether the user shows an affective or emotional response when presenting the information.
- Ability to understand: User's judgement that he/she will be able to understand information presented.
- *Background experience*: Degree of knowledge with which the user approaches information.
- *Content novelty*: The extent to which the information presented is novel to the user.
- *Source novelty*: The extent to which a source of the document (i.e., author, journal) is novel to the user.
- *Document novelty*: The extent to which the document itself is novel to the user.

By understanding relevance criteria usage (e.g. the frequency or distribution of selected criteria), and eventually understanding their relation to user interaction and their effect on relevance judgement, we might be able to determine which criteria to make explicit for what types of users within end-user services, and move towards a more comprehensive evaluation of retrieval system performance that takes the user's cognitive process, interaction and tasks into consideration.

3 Talk-aloud protocols

Talk-aloud protocols are based on the idea that talking aloud while solving a task provides a view of the thoughts as the task-solving process is ongoing [7]. In an IR context, using talk-aloud protocols would provide a researcher with a raw view of the relevance-judgement processes that users go through when searching for the literature. By observing these processes, a researcher can examine them and in turn observe the relevance criteria used within those processes.

After the verbal reports have been collected, they are transcribed and segmented into utterances which are then, in turn, encoded. The granularity of encoding performed on the utterances, if any, will depend on the researchers' needs. In this study (Sect. 5), we encoded utterances using one or more labels from the following encoding:

- Interaction: Any utterance that indicates the participant is performing an operation on/with the system or interacting with it, e.g. reading a document, clicking on a document surrogate, going back a page, etc.
- Intent: Any mention of the participant's intentions regarding the obtained information or regarding their actions, e.g. using a retrieved document to impress their supervisor or initiating a search in the hopes of finding a particular type of information.
- Relevance criteria: Any mention of factors that may affect the participant's choices regarding whether they are to keep or not a document, e.g. if the user picks the document because it is a survey.

Utterances encoded as *interaction* were further encoded either as Navigation (e.g. user interacts with the system by closing a document window, or going back a page), or Reading Aloud (i.e. user interacts with the system by reading a portion of the presented text out loud). Utterances tagged as *relevance criteria* were further encoded using the taxonomy of relevance criteria described in Sect. 2.

The encoding of the utterances into relevance criteria was managed primarily by one of the authors. However, we did validate the process by having one of the others sample utterances from the transcriptions and pass it to a third author for encoding. This resulted in an agreement of approximately 87% between the two authors.

While we cannot be sure that participants had voiced all mental processes, we did make an effort to instruct them to voice every thought passing through their minds, not only those thoughts deemed important, and a training session was implemented to naturalise them to the verbal task.

4 Three aspects of relevance criteria

Studies related to relevance criteria have mostly concentrated on qualitative investigations (e.g. [1,14]), or simple statistics presented in tables (e.g. [15]). Our method, in contrast, aims to provide a more comprehensive view of criteria usage from three perspectives that highlight different types of patterns with respect to system, users and sessions.

First of these is *relevance-criteria profiles*, constructed by aggregating the counts of relevance-criteria usage during the course of a user-search session. It is meant to provide a global profile of relevance criteria weights during the session

with respect to individual users or group of users (e.g. working in related research areas). Second of these is *relevance-judgement complexity*, constructed by considering the number of relevance criteria involved in a single relevance judgement. Final perspective is in *session visualisation* which provides a representation of how the relevance profiles and complexities change throughout the session.

4.1 Relevance-criteria profiles

To build the relevance-criteria profile, the utterances coded as mentions of relevance criteria (Sects. 2 and 3) are grouped at the session level and counted; all mentions of a particular relevance-criterion within the search session contribute to a single count for that criterion. A typical relevance-criteria profile, visualised as a chart, looks like Fig. 3, where the x-axis represents criteria, and the y-axis represents the number of times that criterion has been mentioned in the session. To make the numbers comparable across profiles, we normalise the counts within each profile by dividing by the sum of all criteria mentions:

$$\mathrm{rc}'_i = \frac{\mathrm{rc}_i}{\sum_{j=0}^{N} \mathrm{rc}_j} \tag{1}$$

where rc'_i is the new, normalised, count for relevance criterion i, rc_i is the count for relevance criterion i and N is the total number of relevance criteria (in this article, $N = 15$).

Aggregating profiles, for instance, by participant's affiliation or research experience does not require any special processing. Criterion counts are added by restricting the sums and counts to the group for which the profile is being created.

Profiles can be further compared by means of the Jensen–Shannon (JS) divergence measure [12] for comparing profiles as it is based on the Kullback–Leibler (KL) [11] divergence but is symmetric. The JS divergence considers the KL divergence between p and q under the assumption that if they are similar to each other, then they should both be "close" to their average. As the JS divergence is based on the KL divergence, the smaller the divergence the more similar the two profiles are. Normalised relevance-criteria profiles satisfy the properties of discrete probability functions so they can be compared using this divergence measure.

4.2 Relevance-judgement complexity

While the relevance-criteria profile defined in Sect. 4.1 tells us which criterion features as the most prominent throughout the session, it does not tell us, on average, how many criteria are used for single relevance judgements nor does it illustrate the average complexity throughout a session or across users. To examine this aspect of the relevance-criteria usage, we employ relevance-judgement complexity defined as the number of criteria used in a single relevance judgement.

In this study, we have approximated the criteria involved in a single relevance-judgement process by those in the set of criteria mentions obtained when we delimit the verbal report by utterances that have been encoded as user interaction. That is, we assume the set of criteria mentioned between two interactions (Sect. 3) to represent a single relevance-judgement process. We will refer to a relevance judgement to be of complexity n if the size of the criteria set at that point is n.

The approach of delimiting by interaction steps is very likely to introduce some noise into the analysis, as some judgement processes will span across several interaction steps. However, taking this approach has two immediate benefits: (1) it prevents the introduction of noise coming from subjective annotation of judgement process boundaries, and, (2) it may provide insight into criteria profile changes that invoke interaction steps.

Complexity can be examined as *Polarised Complexity*, where relevance-criterion usage with respect to relevance judgements are counted to reflect whether they have been used as a positive aspect (i.e. implying greater relevance) or a negative aspect (i.e. implying less relevance) of the document.

Consider, as an example, the following excerpt from a potential transcription: "…I'm scrolling down to see the next ten…I think I'm going to click on that topic…yeah, that looks good…that's from 2007 so that's good…oh, but it's only 2 pages long…I know him, I met him at a conference…nah, will close it…." Segmenting and coding the utterances would be achieved as follows. At first, the participant is retrieving the next ten documents in the search results for a combination of topics. This is an interaction with the system and in particular one of navigating (as opposed to reading out loud). Next, the participant decides to click on a link to retrieve the full document. Again, this is a navigation interaction. These two interactions would be encoded as *N*. Next, the participant would mention that the article is 'from 2007,' suggesting that the article is recent enough to warrant more careful evaluation. This utterance would be encoded as *currency*. Next, the participant expresses that the article is too short to be of use (utterance encoded as *depth/scope/specificity* as it references the volume of the information presented). In addition, the participant realises that s/he knows the author (or one of the authors) of the article; however, the final judgement is that the document is not to be kept. These last two utterances are encoded as *quality of sources* and *N*, respectively.

In addition to labelling the utterances, their polarity (whether they have been expressed in a positive or negative fashion) is determined and added to the encoding. The procedure to determine the polarity of any one encoded utterance is described in Sect. 5.2.3.

4.3 Session visualisation

The visualisation technique rests on the 'relevance-criteria piles' metaphor. These piles represent relevance-judgement processes. A relevance-judgement process is then defined as the sequential use of relevance criteria as delimited by interactions.

To plot a search session, first, we group the tagged utterances in relevance-criteria groups. For each group, we plot the first relevance criterion in the sequence at the bottom of the pile, the second on top of it one unit to the right and so on. Blocks are made as long as necessary so that the final shape of the pile resembles a staircase. An example graph can be seen in Fig. 1. The symbol *N* in this figure denotes an interaction step of navigating away from the page and the minus sign refers to a negative mention of the corresponding criterion.

Visualising data using our method can help one uncover emerging patterns in the users' interaction behaviours and relevance-criteria usage, e.g. it can highlight:

- characteristics of anomalous search sessions,
- potentially noteworthy patterns in the order and grouping of relevance criteria occurrences,
- connections between relevance criteria occurrence patterns and subsequent interactions, and,
- the changes that take place to relevance-judgement complexity and -criteria profile as the session progresses.

There are assumptions behind the piles metaphor. First of all, there is the assumption of aggregation. When a relevance criterion has been observed, we assume that this criterion will apply all the way until the user has made a final judgement. The application of criterion is done sequentially until the user is able to make a judgement about the relevance of the information. The length of each block in the graph symbolises this assumption. One of the consequences, should this assumption hold true, is that the sequence in which criteria are used matters and that there might be a degree of relationship between relevance criteria. Users might follow a pattern when using relevance criteria. By using piles, we can start analysing whether a user's relevance-judgement process exhibits these dependencies between relevance criteria. We also assume that each criterion contributes, either negatively or positively, to a final judgement. Negative contributions are represented as a minus sign next to the block in the graph.

Fig. 1 An example with three relevance criteria and interactions plotted

A second assumption is that we can isolate or delimit relevance-judgement processes by the appearance of interactions. We observed that relevance judgements usually end with the user navigating away from the document. This interaction can be preceded by the explicit verbalisation of the relevance judgement, e.g. the user utters 'I don't like this document.' A pile is then defined as occurrences of utterances that are not interactions. There are, however, some shortcomings attached to these assumptions. First of all, depending on what the researcher considers to be an interaction, piles will (or will not) correspond to documents, and their judgement processes as interactions are not necessarily all navigation interactions. Further encoding of interactions might alleviate this to a certain extent, since the dynamics of the session might become more visible. Gathering click-through data and using it to better delimit the relevance-judgement processes might also alleviate this situation.

Plotting sessions using our technique allows a researcher to investigate the relative strength, or importance, of a relevance criterion. Figure 1 corresponds to the example in the previous section. To plot the graph, we first delimit the relevance-judgement process by using interactions as delimiters. This results in the sequence *N N Process N*. Hence, the graph begins with two navigational interactions depicted as *N* each of which correspond to the utterances '…I'm scrolling down to see the next ten…' and 'I think I'm going to click on that topic…yeah, that looks good…', respectively. Next, we have a pile of coloured blocks, each representing a different criterion used in the process. One of the blocks has a negative sign next to it, denoting that the criterion has been used in a negative fashion (negative polarity). In the example, this block corresponds to the utterance '…but it's only 2 pages long…' which has been encoded as *depth/scope/specificity*. Although the criterion has been mentioned in a negative fashion, the judgement process continues. When interpreting the graph, this may suggest that the strength of the criterion, relative to the overall judgement process, is not as strong as to end it right there and then. The explanations can be varied; however, the point is that researchers can direct their attention to further investigate these scenarios. The sequence ends with a navigational interaction.

4.3.1 Choosing a colour sequence

According to Ware [16], the effectiveness of coding using colours for coding is degraded as more categories are added. Ware recommends 12 colours which are normally used when labelling using colours. The first six colours, which also correspond to the basic colours in the colour opponent theory [9], are white, black, red, green, yellow and blue. The remaining six colours are pink, grey, brown, magenta, orange and purple.

Taking the colours as an ordered sequence of recommendations, we use the number of occurrences of relevance criteria, in an aggregated profile, as indices to select an appropriate colour. The most frequently occurring relevance criterion is then assigned the first colour in the sequence, the second most occurring criterion the second colour in the sequence and so on. The rationale behind this procedure is that, since aggregated profiles are obtained by averaging across users, higher relevance criteria counts mean that users have mentioned the criterion, on average, more often; hence, it is likelier to be observed in any one search session. Choosing the most contrasting colours for the most commonly occurring relevance criteria should make easier the visual detection of the different criteria.

5 Results

In this section, we present and discuss how the analytical tools described in earlier sections have been used to analyse the data gathered during a user study carried out from January to August of 2008. A total number of 21 people accepted the invitation to participate in the study. These participants were characterised by three types of affiliation (10 subjects from computing, 8 from information management and 3 from pharmacy). They were further grouped according to their levels of research experience (10 Ph.D. students, 7 researchers and 4 senior researchers) and assigned a task according to this level: writing a literature review for a thesis, framing the impact of a grant proposal, and preparing a keynote speech at a conference, respectively.

The main characteristic of the search task given to users was that it required them to search within several target areas outside their research field for the literature related to their own area of research. To do so, participants were provided with a system designed to return a list of suggested linking topics as well as two sets of documents (see Fig. 2). While the linking topics were placed in a central panel at the top of the screen, the document sets were arranged into a left and right panel below the panel containing the linking topics. The layout of the panels was intended to reflect documents that were more closely related to their own area of research and target area of research, respectively. Full details of the search system employed for this study can be found in [5].

5.1 Comparing relevance profiles

There were 1755 utterances identified as one of the relevance criteria listed in Sect. 2. The global profile, aggregated from all the individual profiles (as defined in Sect. 4.1), is depicted in Fig. 3. We can immediately observe that *tangibility* and *depth/scope/specificity* are the most mentioned criteria (approximately 33.8% and 23% of all criteria mentions, respectively). We also note that *document novelty, affectiveness* and *quality of source* stand next in line (approximately 12.1%, 8%, and 4.82% of all criteria mentions, respectively).

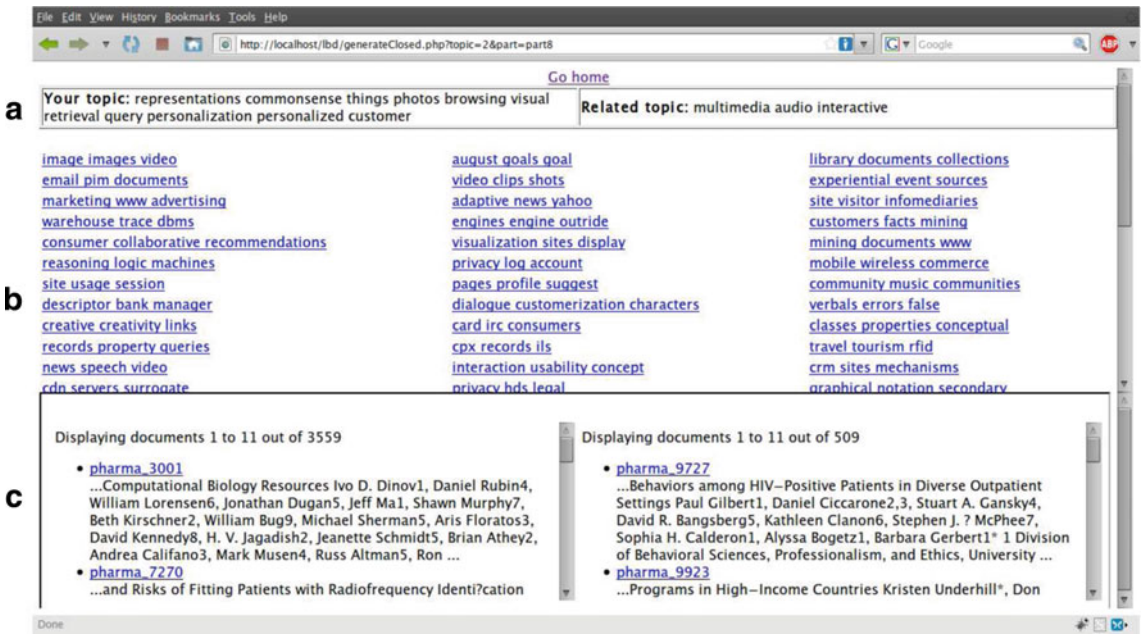

Fig. 2 Layout of the panels in the system used by the participants in the study. The top panel (*a*) contains the start and destination topic. In the middle panel (*b*) we see the linking topics. At the bottom of the screen (*c*), there are two panels, the left and right panel, both of which contain the search results

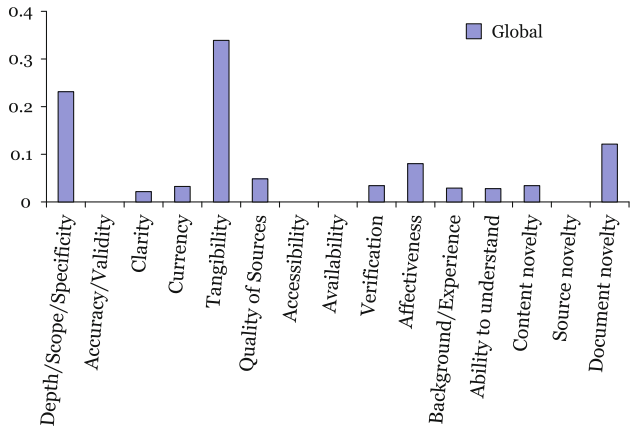

Fig. 3 Global aggregated relevance profile

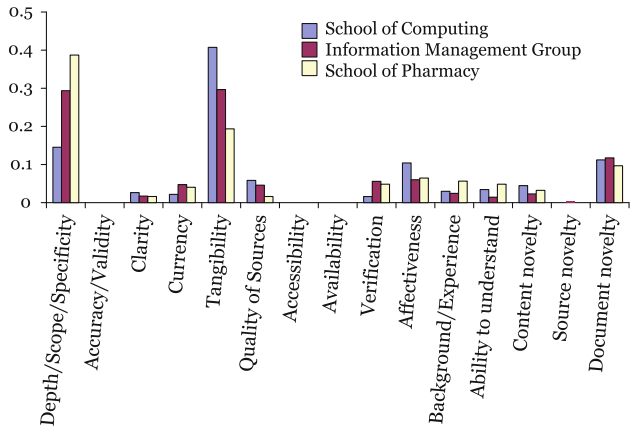

Fig. 4 The school profiles plotted together

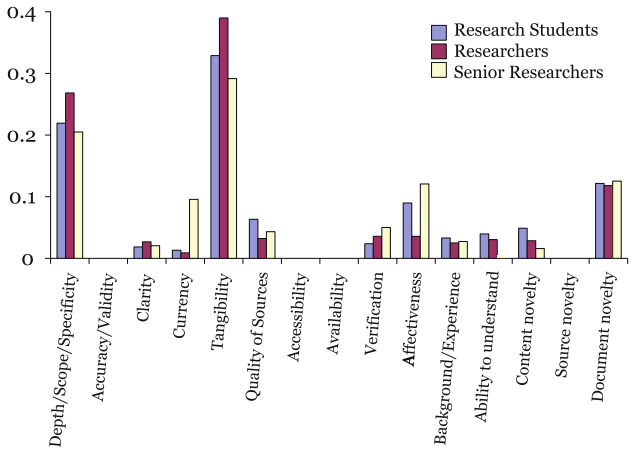

Fig. 5 The profiles according to research experience plotted together

These five criteria make up more than 82% of criteria mentions. All other criteria are comparable at 2–4 per cent of all mentions apart from *accuracy*, *accessibility*, *availability* and *source novelty* which is not mentioned at all.

In Fig. 4, the profiles of the three schools are plotted together. By plotting the profiles together, we can quickly see similarities and differences. In the figure, we observe that while participants from the School of Computing have a distinguishable preference for tangible data, members of the other two schools prefer other aspects of the information such as its depth, scope and specificity. Furthermore, we can also observe that members from all the three schools share the same interest (in terms of proportions) for the novelty of the documents found.

The profiles for groups representing different levels of research experience are plotted in Fig. 5. We immediately observe that researchers and senior researchers seem to apply *affectiveness* (i.e. emotional responses such as whether or not they like the document) more frequently than students. In fact, in the case of senior researchers, there is hardly any difference between mentions of *affectiveness* and that of *document novelty*. Further, senior researchers seem to consider *currency* as an important factor while this does not seem to be a prominent criterion for researchers and students.

By plotting the divergence scores between all participants's profiles and each other, we can spot outliers but also see if there are any naturally emerging groups. The JS divergences between each individual profile and the other profiles are depicted as a matrix in Fig. 6.

In each matrix, the value in cell (i, j) corresponds to the JS-divergence value between the profiles of participants i and j. Rows and columns are ordered by date in which the participant took part of the study. This leads to the participants being ordered by school, i.e. index values from 1 to 10 represent the School of Computing, from 11 to 18 the Information Management Group and from 19 to 21 the School of Pharmacy. The matrices in each map are all equal, and the only difference between maps is the number of colours used as palette for the JS-divergence values; the redder the colour of the cells, the less divergent the two profiles are.

In all matrices, the profile in row/column 6 has a high divergence with almost all the other profiles. This suggests that the participant represented by the profile in row 6 is an outlier. In the last matrix of Fig. 6d, we can observe that the profile in row 18 diverges with practically every other profile but two. One of these two profiles is that in the row 11 which also seems to diverge with most other profiles. In the figure we can also observe that the profiles of the participants of the School of Computing remain fairly convergent and that they diverge more with the profiles of the members of the School of Pharmacy than with those of the Information Management Group. There seems to be a group of profiles that are convergent, to a certain extent, with almost every other profile. These profiles are those in rows 1, 2, 3 and 7 (members of the School of Computing) and 12 and 17 (members of the Information Management Group). That these profiles are convergent with most other profiles could be because the

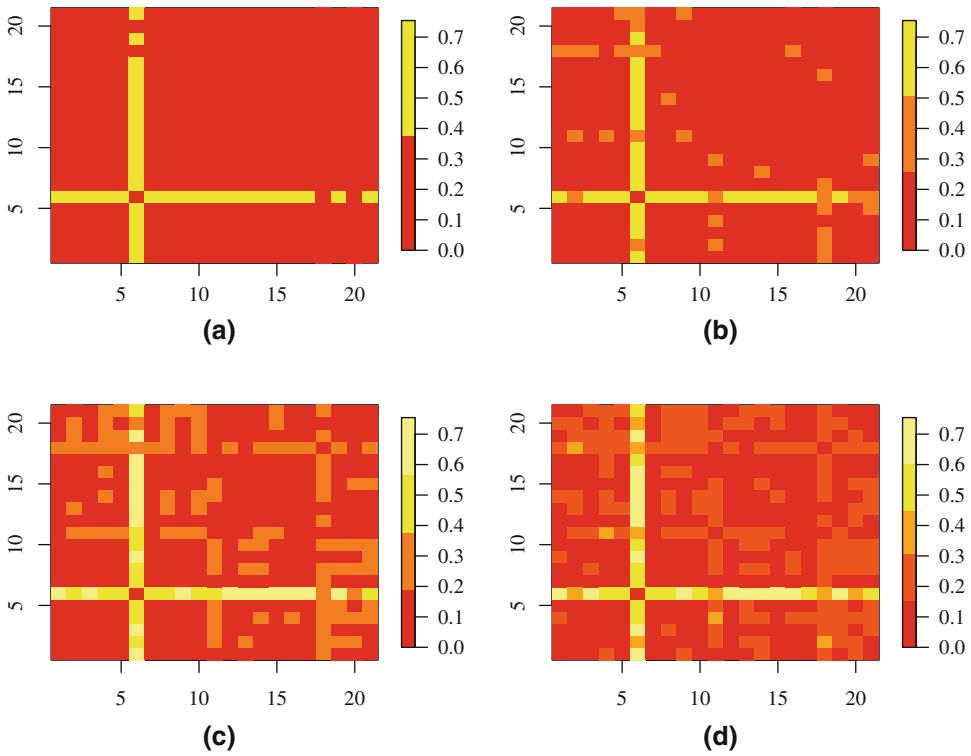

Fig. 6 Jensen–Shannon divergence measure between all individual profiles and the global profile. In each of the four matrices, the (i, j) cell represents the Jensen–Shannon divergence value between the profiles of participants i and j. The colour bar on the left-hand side of each matrix indicates how large a difference has been distinguished by colour in the matrix

participants represented by these profiles follow a globally shared behaviour in using relevance criteria to judge the relevance of the information presented.

5.2 Comparing relevance-judgement complexities

5.2.1 Identifying decision rules

In the study conducted by [15], participants were asked to select, from the results of searches conducted by librarians, which documentation they would use for their projects. One of the observations resulting from the analysis of the selection process is that users applied a set of decision rules when selecting this documentation. The selection process, as described, consisted of six rules:

1. Single criterion decision: If the user detects a single salient unwanted aspect in the information, it is immediately discarded. This rule represents the principle of least effort.
2. Multiple criteria decision: If users cannot reach a judgement after applying the single criterion rule, they apply several criteria until a judgement is reached.
3. Dominance rule: Users select documents such that they excel in at least one criterion and are no worse in any of the other criteria, e.g. two documents which provide the same information; however, one of them is more current than the other.
4. Scarcity rule: When information is scarce, users tend to be more lax regarding the criteria used in judging the information.
5. Abundance rule: When users have found enough information, they tend to stop accepting more information even if it would be deemed relevant under different circumstances.
6. Chain rule: When users have detected that they are on a chain, or vein, of information they tend to make a collective information on the set; for example, because the previous document, deemed relevant, is on this chain, a new document on the same chain is likely to be considered relevant.

It was observed within our study that participants do indeed use some of these rules and that they used them in varying proportions. For example, the dominance rule was mentioned by some participants who assessed the relevance of some documents in relation to the previously assessed documents:

> …that's the kind of paper that I'm looking for, it's probably the most appropriate that I have found, more than previous ones…

> …this must be one of the best ones I've found so far…

The use of the chain rule was also observed:

> …can't help feeling that this one should be a rich vein…

> …that [topic] was quite a rich one so …I got quite a lot out of that one…

One mention was even coupled with a suggestion for a desired feature of the system:

> …this is something that I want, I'm not going to read it because the title says it all …now what I really desperately want is a little box at the bottom that says 'find lots of other things like this'…

That the participant requested a feature that retrieved 'more like this' suggests that the participant suspected that the document might have been the first example of a set of documents in the same vein of information and that they might have all been interesting. One could, therefore, consider this expression as a use of the chain rule.

Most of the rules mentioned above are difficult to quantify without the help of in-depth discourse analysis and extensive human labour. Here, we present an alternative type of analysis which is less intensive and provides an initial overview of the uses of the rules. We exemplify how such analysis can be conducted by doing initial analysis with respect to the first two rules. We estimate the frequency at which the first two of these rules are applied (Sect. 5.2.2) using the relevance-judgement complexity defined in Sect. 4.2. We further estimate how a single criterion might be applied in practice to filter out or eagerly accept a piece of information, by further distinguishing relevance judgements of complexity 1 as polarised complexities reflecting negative and positive applications of relevance criteria (Sect. 5.2.3).

5.2.2 *Relevance-judgement complexity*

A total of 589 uses of selection rules (of any complexity) were counted, where the application of a selection rule is taken to be any relevance-judgement process found within the verbal report (see Sect. 4.2 for a description of how relevance-judgement processes are obtained). Out of these, 215 (36.5%) correspond to relevance judgements of complexity 1, whereby a single criterion has been used in making a decision on the information evaluated (corresponding to the single criterion rule) and 374 (63.5%) to relevance judgements of complexity $n \geq 2$ (corresponding to the multiple criteria selection rule).

The bars in Fig. 7 display the total number of participants that used a sequence of n criteria to assess the relevance of the information presented. We observe that all participants applied, at least once, relevance judgement of complexity 1 to the information presented, i.e. all participants based their judgement of the information presented using a single criterion at least once during their session. We also observe that the majority of participants (14 participants) used up to six criteria in any one relevance-judgement process. More

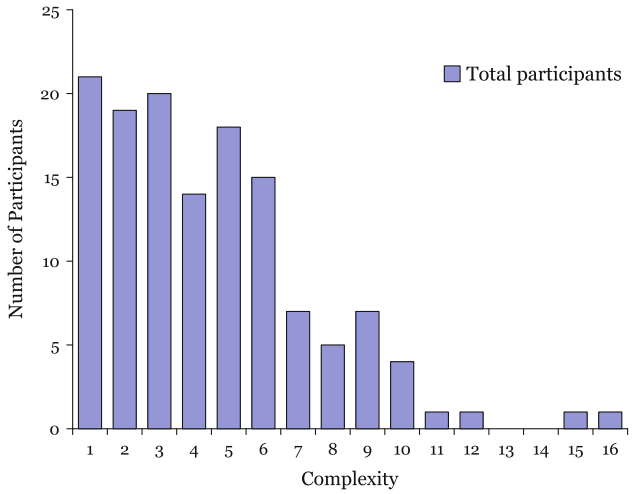

Fig. 7 Total number of participants (Y-axis) that used, at least once, a relevance-judgement process of complexity n (X-axis)

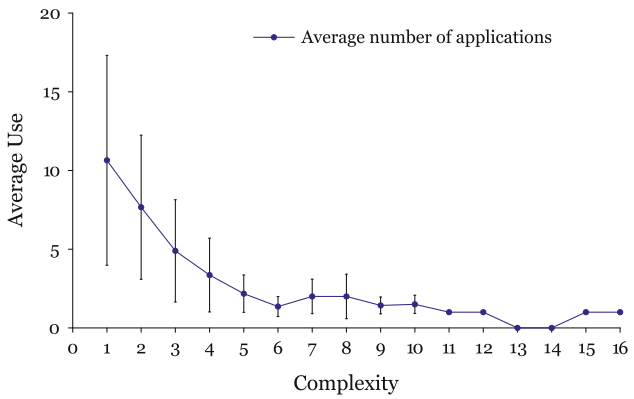

Fig. 8 Average use (Y-axis) of relevance-judgement processes of length n (X-axis). *Bars* represent a standard deviation

complex relevance-judgement processes seem to be used by fewer participants. Processes of complexity larger than 7 were used by, at the most, seven participants.

In Fig. 8, we have presented the average number of relevance judgements of complexity n per search session. In the figure, we observe that sequences of complexity 1 (single criterion rule) were used, on average, about 10.6 times per session. As seen in Table 1 (as well as in Fig. 8), the average number of uses of relevance judgements of complexity $n \geq 2$ is always lower than the average use of relevance-judgement processes of complexity 1. In addition, in this study, we found a correlation (0.01 level at p value -0.923) between complexity and average number of uses: we expect to observe a greater number of less complex relevance-judgement processes to be used on average (and by more participants) than more complex relevance-judgement processes. This does not imply that these processes are less complicated, just that fewer criteria have been used (see Sect. 4.2, for our definition of complexity).

Table 1 Average use (averaged across participants that expressed using them) of relevance-judgement processes of complexity *n*

Complexity	Average use (times)	Deviation
1	10.65	6.6
2	7.6	4.5
3	4.9	3.2
4	3.35	2.3
5	2.17	1.1
6	1.35	0.6
7	2	1
8	2	1.4
9	1.42	0.5
10	1.5	0.5
11	1	–
12	1	–
13	–	–
14	–	–
15	1	–
16	1	–

Table 2 Distribution of each criterion as distributed across single criterion rule uses

Criterion	Single (%)	Multiple (%)
Depth/scope/specificity	6.4	93.6
Accuracy/validity	–	–
Clarity	7.9	92.1
Currency	15.8	84.2
Tangibility	11.9	88.1
Quality of sources	8.2	91.8
Accessibility	–	–
Availability	–	–
Verification	8.3	91.7
Affectiveness	12.8	87.2
Background knowledge	9.8	90.2
Ability to understand	20.4	79.6
Content novelty	8.3	91.7
Source novelty	–	–
Document novelty	26.8	73.2

In Table 2, we have presented the percentages of mentions with respect to each criterion (first column) which have been used in single criterion selection rules (second column) and multiple criteria selection rules (third column). The figures show that, while there are criteria, such as *document novelty*, which seem to have a stronger tendency to be used in a single criterion rule (26.8% of all mentions of *document novelty* are in a single criterion selection rule), all of the criteria have been used at least 6.4% of the time in a single criterion selection rule. That is, there was no criterion that could be regarded as a criterion that does not play a role in immediate dismissal or acceptance of a document.

5.2.3 How information is filtered: polarity in criteria usage

The single criterion decision rule, as described by [15], suggests that this rule is mostly applied to quickly dismiss information based on salient unwanted features. During this study, however, two types of use of the single criterion rule were observed:

- Filter out: In concordance with the original description of the rule, participants detected salient unwanted features and quickly dismissed the information.
- Eager acceptance: Contrary to the original mention of the rule, participants detected a salient feature that made them consider the presented information relevant automatically.

The frequency with which these two uses were observed can be estimated as follows. Firstly, we counted the use of each criterion in a single criterion rule with respect to its polarity. A negative mention of *currency*, for instance, was considered different from a positive expression of the same criterion. By making this differentiation, we assume that there is a concordance between negative mentions of criteria and uses of the rule to filter out information considered irrelevant and between positive mentions of criteria and uses of the rule to eagerly accept the relevance of the presented information.

The assessment of the polarity of any one utterance was done by analysing the type of words used in the utterance itself. In the few cases where the language itself was not enough to determine the polarity, the tone of the voice of the participant and the preceding utterances were taken into account.

Consider the following example. A participant mentions that '…[the document] is too old…'. This utterance is classified as *currency* and its polarity deemed negative. The negative polarity is inferred from the use of 'too old' in the utterance. This expression suggests that the participant deemed the information to not fulfil a specific criterion: that the information is current or up to date. *Currency* is used as a criterion, but in a negative fashion and will probably influence the final relevance judgement of the information presented as being negative.

On the other hand, consider the polarity of utterances such as '…it's from 2006…'. The language used in the utterance indicates that the person is referring to the date of publishing and may be referring to the potential *currency*

Table 3 Frequency of each criterion as distributed across single criterion rule uses

Criterion	Positive		Negative		Total	
Depth/scope/specificity	9	4.1%	17	7.9%	26	12.0%
Accuracy/validity	–	–	–	–	–	–
Clarity	3	1.3%	–	–	3	1.3%
Currency	3	1.3%	6	2.8%	9	4.1%
Tangibility	**55**	**25.5**%	16	7.5%	71	33.0%
Quality of sources	7	3.3%	–	–	7	3.3%
Accessibility	–	–	–	–	–	–
Availability	–	–	–	–	–	–
Verification	2	1.0%	3	1.3%	5	2.3%
Affectiveness	13	6.0%	5	2.3%	18	8.3%
Background knowledge	3	1.3%	2	1.0%	5	2.3%
Ability to understand	1	0.4	9	4.2%	10	4.6%
Content novelty	3	1.3%	2	1.0%	5	2.3%
Source novelty	–	–	–	–	–	–
Document novelty	16	7.5%	**41**	**19.0**%	57	26.5%
Total	114	53%	101	47%	215	100.0%

The percentages are calculated across all the 215 single criterion judgements

of the information; however, it does not offer any indications regarding the polarity of the expression. In cases like this, the audio recordings were used for assessing the tone of the person's voice along with an analysis of the preceding utterances. For example, suppose that the first mention of a criterion is negative and that it is to be encoded with *depth/scope/specificity*, e.g. '…it's only 2 pages long…'. As expressed, the user is already starting to lean towards a negative judgement. Should the next utterance be '…[and] it's from 2006…', then its polarity would be deemed negative. This stems from the use of the word *and* to connect the two mentions of criteria suggesting that they share the same polarity. On the contrary, should the next utterance be '…[but] it's from 2006…', then its polarity would be deemed positive. In this case, what makes the polarity to be deemed positive, instead of negative as in the first example, is that the expression is contraposed by the appearance of the word *but* which signals an opposite polarity to that of the first utterance (negative). The preceding utterance and its polarity are used as a reference point against which the polarity of the following utterance is judged.

The counts depicted in Table 3 show that criteria mentioned are almost evenly distributed across polarity; out of a total of 215 criteria mentioned, 114 correspond to positive mentions while 101 are negative mentions. All criteria were used—either positively, negatively or both—at least once, in a single criterion rule.

Because the verbal data gathered from participants did not always correspond to actual relevance judgements of documents, some of the uses of the single criterion rule were observed in a different context. Positive uses of this rule were used mostly for assessing the *potential* relevance of the information. That is, participants expressed using a criterion in a positive fashion to decide whether the information could be relevant. The relevance of the information would then be decided, possibly by using more than one relevance criterion, once it had been assessed more thoroughly. Negative uses, on the other hand, were always used for immediately dismissing the information and hence corresponded to negative judgements of relevance.

Filtering out irrelevant information was mostly done on the grounds that the documents were not novel, e.g. a document had been re-retrieved. Participants mentioned *document novelty* in a negative fashion 41 times (about 19%) when using a single criterion rule:

> old ants! ah the little buddies …that's the document I already have now so I'm not going to read that …
>
> …yes, I've seen it before …oh, not again, no, still not want to see that, hmm …

The second (-) most used criterion, for filtering out irrelevant information was*depth/scope/specificity*. Document length, in particular, seemed to be an important factor when assessing the relevance of the information:

> …no it's very short, I'll put it back …
>
> …I'm gonna put it back because it's very brief and a bit journalistic …

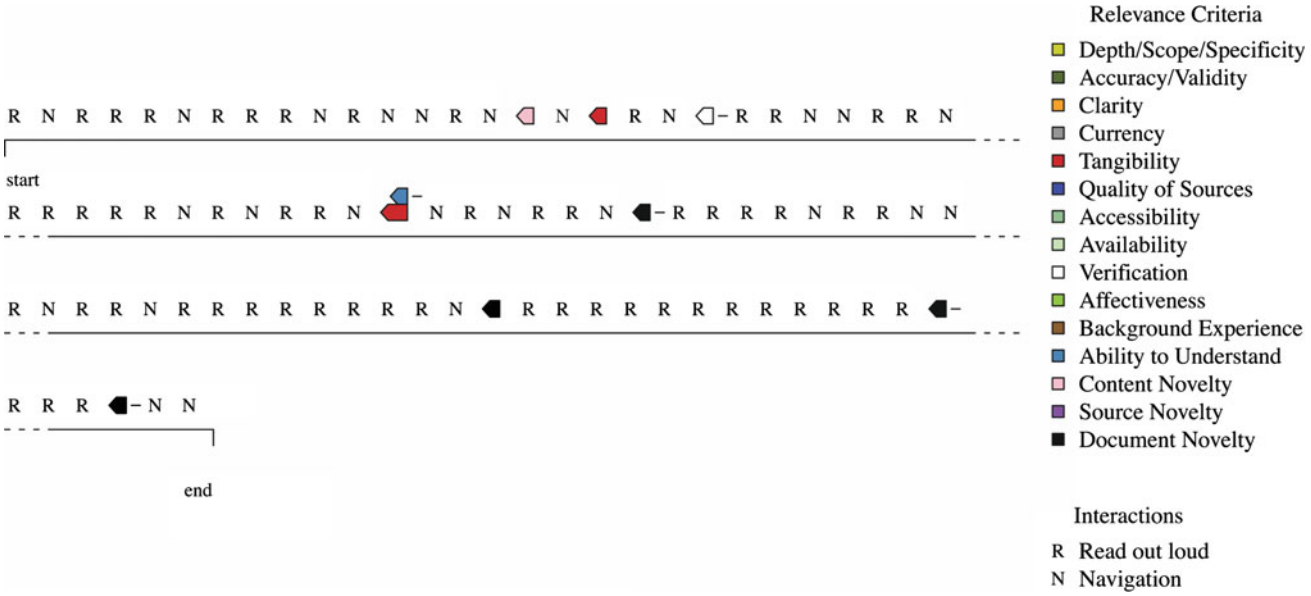

Fig. 9 The anatomy of an anomalous search session

The most used criterion in positive relevance judgements made using the single criterion rule is *tangibility* as the most used criterion. This may suggest that some participants found hard data a good indicator of the relevance of the information and when this criterion was met, they were quick to accept the information as relevant. However, one must remember that mentions of topicality were also encoded as *tangibility*:

> …oh yes it's about simulations, interactive kind of thing, I'll write that one down as well …

> hmm, yeah, that could be an interesting application, all right, oh I need to write this stuff at the top don't I?[2]

As observed earlier, *document novelty*, when used negatively, seems to be an indicator of irrelevance. This observation, coupled with that *document novelty* is the second most used criterion in single criterion relevance judgements suggests that the correlation between relevance judgements and the polarity of *document novelty* may be high. *Document novelty* was mentioned in positive judgements as

> …again that one has already been identified as high up which is really emphasising to me that I probably should read it first, and it probably is a major one in this I'm kinda liking this one …

> …I think maybe I've seen before, again it gives me a lot of theoretical underpinning it has a lot of really nice, well not really nice, mathematical stuff anyway and yeah I think that's probably the one I would take …

5.3 Plotting sessions in practice

The visualisation technique we introduced in Sect. 4.3 can provide an effective means of grasping session characteristics with respect to relevance-criteria profiles, judgement complexities and information-filtering behaviour. It can quickly assess the potential of the information system as well. By incorporating a new dimension, that of time, the visualisation technique provides a more detailed view of the dynamics between interactions and judgement processes as well as the dynamics between criteria within judgement processes.

For example, a much quicker approach to confirming the anomalous behaviour of the diverging profile of participant 7 found in Fig. 6 would have been to look at the visual representation of the participant's search session. This visualisation is presented in Fig. 9. At first sight, it can be seen that the participant not only did not mention relevance criteria very often but also that the participant spent almost all of the session reading out loud. This could reflect a misunderstanding in the instructions for the study or simply that the participant found no document that was even remotely interesting.

In fact, the session consists of eight low complexity piles, and five out of the eight piles end with a negative mention of a relevance criteria followed by an interaction step. Three of these negative mentions is in relation to *document novelty*. Evidence was provided in Sect. 5.2.3 that this criterion was often used negatively to filter out information. This perhaps suggests that the search system did not produce appropriate documents for the participant.

[2] In the study, participants were asked to write down the document identifier whenever they thought they wanted to keep it for later reference, i.e. they considered it *relevant.*

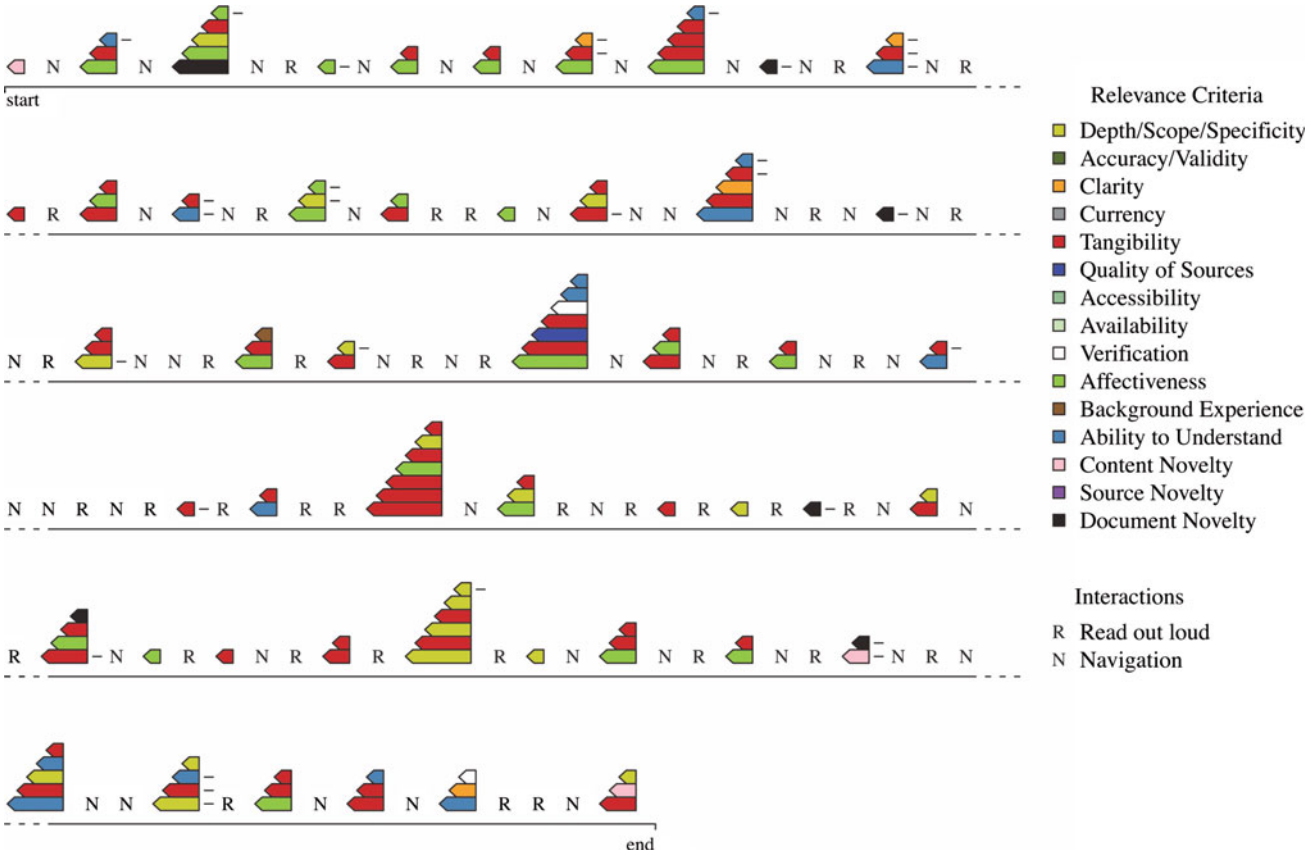

Fig. 10 Visual representation of the search session of participant 2

The participant in Fig. 10 is a research student from the School of Computing. At a glance, if we interpret the number of expressions of *affectiveness* as a measure of engagement, then we can observe that the participant is engaged from the beginning, and remains so throughout the session. These affective responses are represented as blocks coloured in light green. Effectively, out of 49 relevance-judgement processes (depicted as coloured piles in the graph), 22 (about 45%) contain at least one expression of *affectiveness*. Affective responses seem to be, however, more frequent at the beginning than when closer to the end of the session. Perhaps the participant begins to express less emotions (or have less emotional responses) as the session progresses, and s/he becomes more familiar with the underlying collection.

In addition, *tangibility*, which includes topicality, seems to play an important role during the participant's search session. Out of the 49 relevance-judgement processes, 37 (about 75%) include at least one utterance encoded as *tangibility*. This complements the global view presented by the relevance-criteria profile (see Fig. 4) which showed that *tangibility* was a commonly used criterion by participants from the School of Computing. During the participant's session, *tangibility* not only was a commonly used criterion, but also one that was present in most relevance-judgement processes. Moreover, the criterion is present in relevance-judgement processes of different complexities covering almost the full range.

The participant seems to engage in simple to semi-complex relevance-judgement processes very often. The interweaving of piles and interactions (including acts of reading out loud) is frequent. This may suggest a more 'careful' approach at searching for relevant information. A frequent alternation between interactions and uses of relevance criteria may be due to the participant constantly analysing the presented information looking for cues to derive its relevance. As such, it may be a sign of the participant's experience in finding these cues. A person relatively inexperienced in finding these cues may have to sequentially assess each information piece in more detail. This would be translated to stacks of two or three relevance criteria blocks. It may also be that the participant is wary and does not want to filter out potentially relevant information too quickly. Hence, the participant assesses in more detail (than average) each piece of information. As the participant expressed: '…hmmm …I'm usually crap at selecting things for my literature review, I either go for everything or select hardly anything …' which suggests that the participant will use a more careful strategy.

The participant in Fig. 11 is a senior researcher from the Information Management Group. In the researcher's search

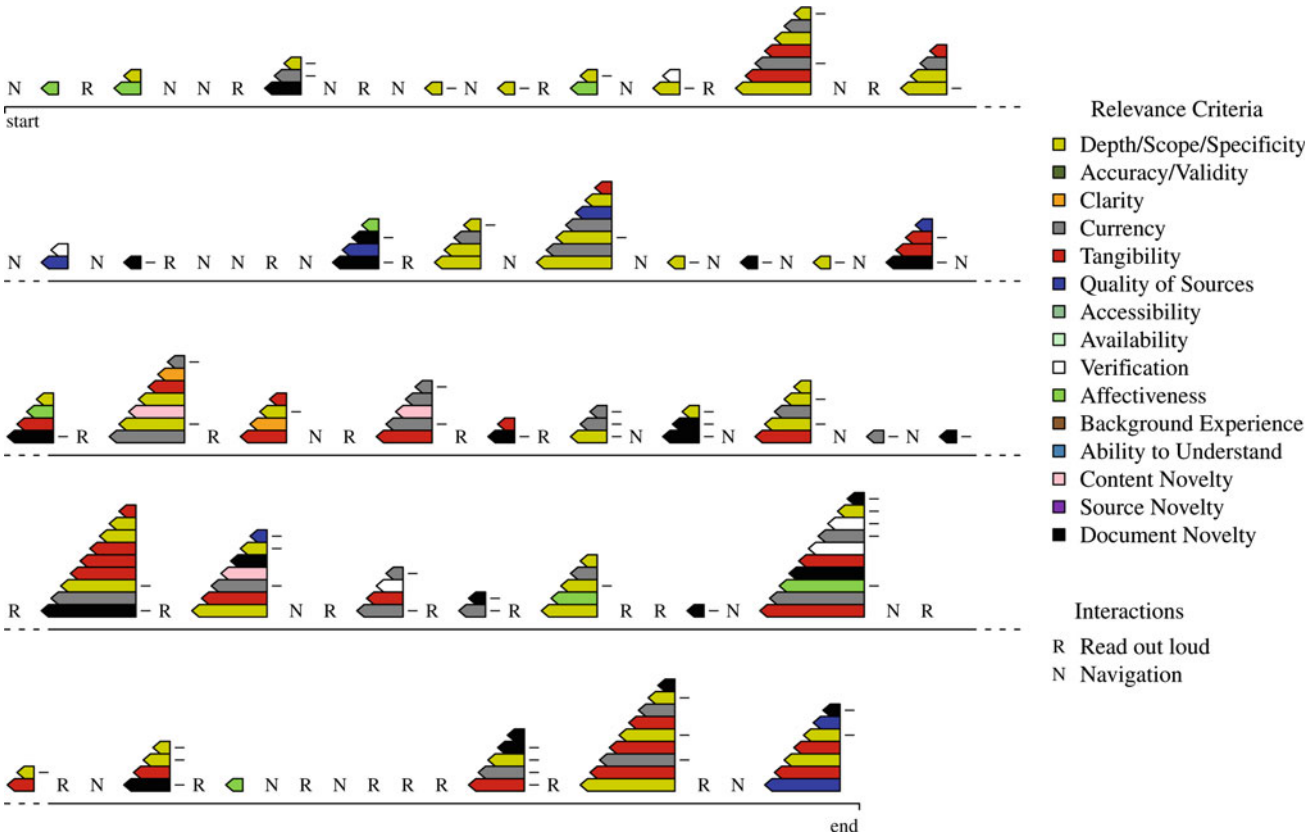

Fig. 11 Visual representation of the search session of participant 19

session, we can observe that *tangibility* is not as prominent a criterion as it is for the participant in Fig. 10. Effectively, out of 41 relevance-judgement processes, 19 (about 46%) contain at least one use of *tangibility* as relevance criteria. *Depth/scope/specificity*, on the other hand, appears at least once in 27 (about 65%) relevance-judgement processes. As depicted in Fig. 4, members of the Information Management Group mentioned in near-equal proportions the criteria *tangibility* and *depth/scope/specificity*. The participant in Fig. 11, however, seems to unbalance this proportion in favour of *depth/scope/specificity*.

Contrary to the interaction behaviour exhibited by the participant in Fig. 10, the participant in Fig. 11 seems to navigate more diligently. Whenever the participant considers to have found a promising source of information, however, the relevance-judgement processes are rich both in the number of uses of relevance criteria and in their variety. On average, the relevance-judgement processes in which the participant engaged seems to be more complex than those of the participant in Fig. 10. In Fig. 12, we include a bar chart depicting the frequency of the relevance-judgement processes, which both participants incurred. The participant in Fig. 10 seems to mostly engage in processes of complexity 1, 2 and 3 with some occasions in which more complex processes are used. The participant in Fig. 11, on the other hand, seems to make use, on average, of more complex processes. Even though simple processes (of complexity 1) are used frequently—possibly for quickly filtering irrelevant information—the remaining processes are more evenly 'spread out', and more complex processes are more frequent.

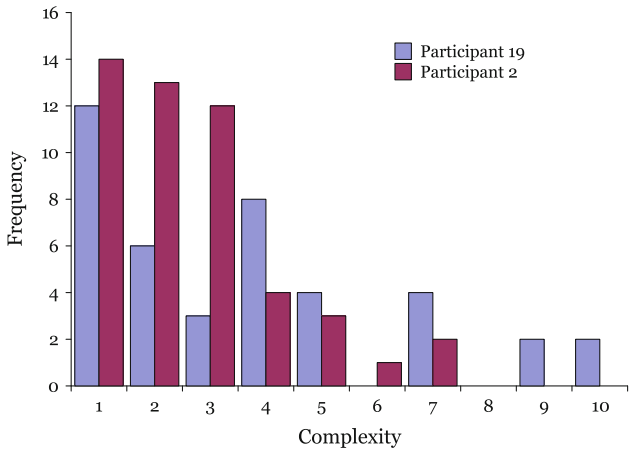

Fig. 12 Frequency of the relevance-judgement processes by complexity

In both the sessions, we observe that some criteria are repeated within relevance-judgement processes. *Tangibility*, for instance, is mentioned up to five times within one relevance-judgement process (the participant in Fig. 10). This is, however, reasonable. The code *tangibility*, as it was interpreted in this study, includes mentions of topicality. Furthermore, several different expressions of references to hard data are to be encoded as tangibility. Expressions like '…[a] neural network, ah you know what that could almost be an application if a neural network can do it you would be able to evolve it as well …' and '…it'd be one to look at to get references from this …' are both encoded as *tangibility*; however, they both refer to different types of tangible information being analysed. One refers to the details of an implementation of a technique (a neural network), and in some sense it could also be encoded as *depth/scope/specificity*, while the other expression refers to the references to be extracted from the document (which could also be encoded as *intent*). *Depth/scope/specificity* is another such code. It was mentioned up to four times within any one relevance-judgement process (the participant in Fig. 10). As a criterion that encompasses mentions of different properties of the information being assessed (its depth, its scope, its specificity with respect to the user's information needs, etc.), *depth/scope/specificity* is likely to be repeated within relevance-judgement processes. Consider these expressions from the participant in Fig. 11:

> …[this] is really what I'm interested in and again is really relevant to the brief which is find new technologies or technologies used in a new way for knowledge management and sharing …looks quite current, november 07, looks a wee bit anecdotal and it's very short so I'll put it back …

These expressions correspond to Fig. 13. As observed in the figure, there are repetitions of *depth/scope/specificity* and even with opposite polarity. This is due to that there are expressions referring to the specificity with respect to the participant's information needs ('it's really relevant to the brief') and to the volume of the information ('it's very short'). This repetition of mentions of the criterion *depth/scope/specificity* was observed frequently for the participant in Fig. 11 (and other members of the Information Management Group), but not for the participant in Fig. 10 (and the remaining members of the School of Computing), while repetitions of mentions of *tangibility* were more frequently observed during the session of the participant in Fig. 10 (and remaining members of the School of Computing).

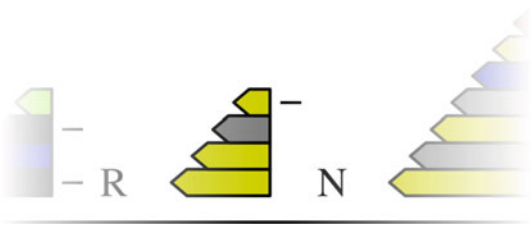

Fig. 13 Repeated expressions of *depth/scope/specificity*

6 Discussion

In this article, we presented the notion of relevance-criteria profiles, relevance-judgement complexity, and a visualisation technique to plot the interactions and relevance-criteria mentions observed during search sessions.

We demonstrated, by example, how these tools aid the analysis of data. First, we showed how aggregated relevance-criteria profiles provide global views of different user groups' preferences. More specifically, we showed how plotting relevance-criteria profiles together can help uncover both (dis)similarities in relevance criteria usage at a global level. For example, outlier detection as well as cluster analysis are two of the types of analysis that can be performed when J(-)S divergence scores between pairs of profiles are plotted together. Second, we discussed relevance-judgement complexity and polarised complexity to show how these can be used for helping us understand user's information-filtering behaviour with respect to different relevance criteria. And, third, we put the visualisation technique presented in Sect. 4.3 into practice to illustrate its potential in aiding the analysis of search sessions. The visualisation reconfirms user's preference for selected criteria observed in their relevance-criteria profiles and immediately highlights levels of relevance-judgement complexity and polarity with respect to selected user sessions.

Session visualisation also offered us the possibility of observing how criteria are repeated *within* relevance-judgement processes. This complements the view offered by relevance-criteria profiles which offer a global view of criteria occurrences. For example, we observed that *tangibility* and *depth/scope/specificity* are repeated in the processes of the participant represented in Figs. 10 and 11, respectively.

One must interpret this observation carefully, however, as the granularities of both criteria are coarse, e.g. mentions of the volume of the information being judged, the specificity and the level of detail of it are all to be encoded as *depth/scope/specificity*. In fact, there are several points to reconsider in a more extensive study before the analysis can be considered conclusive. For example,

– All participants were researcher students or researchers and examined within the context of three tasks only. This may have introduced a bias in the results.

- The classification of utterances was only briefly validated by two of the researchers. More researchers may required for the establishment of some level of consensus.
- The number of people in each group (research background or research experience) may be not be balanced enough for a fair comparison.
- The visualisation is compared across only three participants, making it hard to judge whether the tool would be scalable.
- The assumptions governing the coding and interpretations may be open to debate, such as the theory of sequential application of relevance criteria, and the restriction of relevance judgments to one interaction (when the consideration of a document may well reach across several interaction). Nevertheless, we agreed that this was not an unreasonable first approximation.
- The choice of relevance criteria might need some refinement. For example, 'topicality' is not included as an explicit independent criteria. It is included as a subclass of 'tangibility' and 'depth/scope/specificity' criteria, which poses a problem in the study's comparability to other studies dependent on topicality. Also, the 'novelty' criterion is used for signifying both 'seen before' in general and 'seen before in the session,' which are very different aspects and may cause confusion in analysis.
- The effects of the search interface on the results of the study was not sufficiently analysed.

The last two points, may be weaknesses in the methodology requiring improvement. The other shortcomings listed above with respect to the pilot study is a direct result of limited resources, as the study was conducted as part of a Ph.D. research project. We hope to take these points on board in the next stage of the research.

Relevance criteria are not theoretical concepts, but tangible and concrete. Operationalising them can potentially impact positively on search services, by embedding the most observed criteria explicitly into the system. The criteria associated to selected items can be modelled, potentially,to improve the system's performance in returning relevant information. More immediately, however, the explicit display of criteria related characteristics will enable the user to quickly make decisions about returned information.

Tangibility, may be approximated, for instance, by looking at the number of tables in a document, and *depth/scope/specificity*, by looking at the number of pages in a document (document length has been mentioned frequently as a relevance criteria). Relevance processes, and the intertwined interactions, may be used for modelling user-search behaviours in an attempt to personalise and adapt the system to better accommodate the current information needs of users.

We would like to conclude by observing that the relevance criteria discussed in this article not only have the potential to provide the context necessary for discerning the relevance of information, but also, to provide a basis for understanding the information within the wider context of information space. The criteria clearly maps to contextual elements, such as date and time (e.g. related to currency), people (e.g. related to source, reliability and verification), topicality (related scope) and genre (e.g. research article and/or survey as a proxy for judging depth). The latter set of document characteristics have been highlighted in other studies as significant properties in information understanding (e.g. [8]).

Acknowledgment Supported by the Engineering and Physical Sciences Research Council as part of the project 'Automatic Adaptation of Knowledge Structures for Assisted Information Seeking (AutoAdapt)' [EP/F035705/1].

References

1. Barry, C.L.: User-defined relevance criteria: an exploratory study. J. Am. Soc. Inform. Sci. **45**(3), 149–159 (1994)
2. Barry, C.L., Schamber, L.: Users' criteria for relevance evaluation: a cross-situational comparison. Inform. Process. Manage. **34**(2–3), 219–236 (1998)
3. Borlund, P.: The concept of relevance in IR. J. Am. Soc. Inform. Sci. Technol. **54**(10), 913–925 (2003)
4. Borlund, P.: The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. Inform. Res. **8**(3), (2003)
5. Cervino Beresi, U.: Related scientific information: a study on user-defined relevance. Ph.D. thesis (2010)
6. Cleverdon, C.W., Mills, J., Keen, E.M.: Factors determining the performance of indexing systems, vol. 1: design, vol. 2: test results. In: Aslib Cranfield Research Project, Cranfield (1966)
7. Ericsson, K.A., Simon, H.A.: Protocol analysis: verbal reports as data. MIT Press, Cambridge, MA (1993)
8. Mayer, R., Rauber, A.: Establishing context of digital objects' creation, content and usage. In: Proceedings of the First International Workshop on Innovation in Digital Preservation, Austin, TX (2009)
9. Hurvich, L.M., Jameson, D.: An opponent-process theory of color vision. Psychol. Rev. **64**, 384–404 (1957)
10. Kelly, D.: Measuring online information seeking context, part 2: findings and discussion. J. Am. Soc. Inform. Sci. Technol. **57**(14), 1862–1874 (2006)
11. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 79–86 (1951)
12. Lin, J.: Divergence measures based on the shannon entropy. IEEE Trans. Inform. Theory **37**, 145–151 (1991)
13. Savolainen, R.: The sense-making theory: reviewing the interests of a user-centered approach to information seeking and use. Inform. Process. Manage. **29**, 13–28 (1993)
14. Schamber, L.: Users' criteria for evaluation in a multimedia environment. In: Proceedings of the 54 Annual Meeting of the American Society for Information Science, vol. 28, pp. 126–133 (1991)
15. Wang, P., White, M. D.: A cognitive model of document use during a research project. Study II. Decisions at the reading and citing stages. J. Am. Soc. Inform. Sci. **50**(2), 98–114 (1999)
16. Ware, C.: Color sequences for univariate maps: theory, experiments and principles. IEEE Comput. Graph. Appl. **8**(5), 41–49 (1988)